Learning in Neural Networks Based on a Generalized Fluctuation Theorem

Takashi Hayakawa[*]

RIKEN Brain Science Institute,
2-1 Hirosawa, Wako, Saitama 351-0198, Japan

Toshio Aoyagi
Graduate School of Informatics, Kyoto University,
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
(Dated: April 14, 2015)

Information maximization has been investigated as a possible mechanism of learning governing the self-organization that occurs within the neural systems of animals. Within the general context of models of neural systems bidirectionally interacting with environments, however, the role of information maximization remains to be elucidated. For bidirectionally interacting physical systems, universal laws describing the fluctuation they exhibit and the information they possess have recently been discovered. These laws are termed *fluctuation theorems*. In the present study, we formulate a theory of learning in neural networks bidirectionally interacting with environments based on the principle of information maximization. Our formulation begins with the introduction of a generalized fluctuation theorem, employing an interpretation appropriate for the present application, which differs from the original thermodynamic interpretation. We analytically and numerically demonstrate that the learning mechanism presented in our theory allows neural networks to efficiently explore their environments and optimally encode information about them.

Introduction : The neural systems of animals are prominent as highly efficient systems for processing information concerning the external environment. Many authors have argued that the learning capability of neural systems is information-theoretically optimal by showing that several features of neural activity can be accounted for by positing information maximization (Infomax) in the learning of sensory signals and intrinsic dynamics in neural circuits [1–5]. However, Infomax has not yet been investigated in a general context. In particular, although a real neural system interacts bidirectionally with its environment, not only receiving sensory signals and organizing intrinsic dynamics accordingly, but also generating motor outputs that influence the environment, Infomax learning has not been clearly formulated in this context (see Fig.1(a)). For example, the formulation of Infomax must be generalized in order to facilitate its application to the following type of model. One of the standard models of the interaction between neural systems and environments employs the Markov decision process. In this model, we consider discrete-time ($t \in \mathbf{Z}$) stationary Markovian dynamics of a stochastic neural network with N binary-valued neurons $x^t \in \{0,1\}^N$ interacting with an environment y^t that takes values in a discrete state space $\mathcal{Y}$ [6–8]. The neural elements $\{x_i^t\}$ ($1 \le i \le N$) receive inputs from the environment in such a manner that x^t realizes values stochastically according to a conditional probability $\pi(x^t|y^t)$. This conditional probability depends on the model parameters, such as the synaptic strength, and it changes slowly during the learning process through the adjustment of the model parameters. Then, the state of
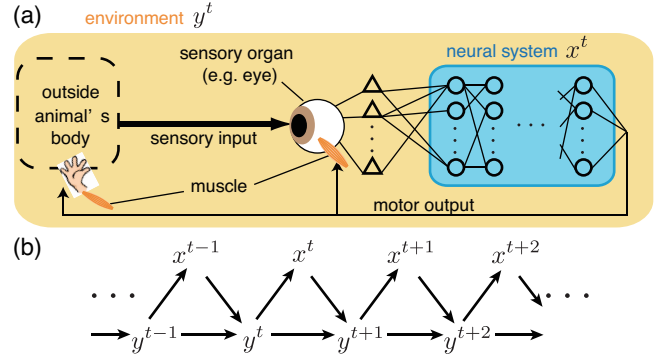

FIG. 1. (Color). (a) Interactions between a neural system and its environment, (b) Representation of the dynamics as a causal network.

the environment at the next timestep, y^{t+1}, is obtained stochastically with a transition probability $\mu(y^{t+1}|y^t, x^t)$ that is determined by the current states of the neural network and environment.

For interacting physical systems, there are recently discovered universal laws called *fluctuation theorems* [9–14] that relate nonequilibrium physical quantities to informational quantities such as mutual information. In particular, the dynamics of the neural network and the environment mentioned above are represented in the form of a causal network with regard to which a generalized version of the fluctuation theorem has been investigated [14] (Fig.1(b)). Because fluctuation theorems describe informational quantities for interacting systems, it is natural to hypothesize that

they may provide a description of a key aspect of the learning behavior exhibited by neural systems. Although information thermodynamic considerations have been investigated in the context of learning systems in a few pioneering studies [15, 16], it has not been determined whether informational quantities are actually maximized in such systems in some systematic way. In this Letter, we study this question.

We derive a novel type of Infomax learning, starting from the version of the integral fluctuation theorem presented in [14], which provides the following inequality relating the average entropy production $E[\sigma]$ of the neural network and the transfer entropy $I_{x \to y}$ from the neural network to its environment:

$$I_{x \to y} \geq -E[\sigma], \quad \sigma \equiv \log \frac{\pi(x^{t+1}|y^{t+1})}{\pi(x^t|y^{t+1})}. \tag{1}$$

Throughout this article, the expectation value E is taken with respect to the stationary distribution p_s of the dynamics unless otherwise noted. The transfer entropy is defined as the conditional mutual information [17]:

$$I_{x \to y} \equiv I[y^{t+1}; x^t | y^t]. \tag{2}$$

As in information theory [18], the (conditional) mutual information between two variables is defined as the change in the (conditional) entropy of one of the two variables owing to the inclusion of the other variable as a conditioning variable. Explicitly, we have $I[y^{t+1}; x^t|y^t] = H[y^{t+1}|y^t] - H[y^{t+1}|x^t, y^t]$. The (conditional) entropy is defined in terms of the stationary distribution as $H[y^{t+1}|y^t] = -E[\log p_s(y^{t+1}|y^t)]$. The quantity $I_{x \to y}$ represents the amount of information that the neural system possesses about the future state of the environment. Thus, from the point of view of Infomax, it is a reasonable hypothesis that maximizing this quantity is an effective learning mechanism. However, it is necessary for the calculation of $I_{x \to y}$ to directly estimate the transition probability of the environment, μ, and this estimation apparently cannot be carried out by the neural network itself. Its lower bound, $-E[\sigma]$, on the other hand, can be computed within the neural network, because σ is determined by the transition probability of the neural network, π, alone. With these in mind, it is natural to conjecture that neural systems attempt to optimize their acquisition of information about the future by adjusting π in such a manner to maximize the quantity $-E[\sigma]$. However, note that the equality in the relation, $I_{x \to y} \geq -E[\sigma]$, is not generally realized, and thus the maximization of $-E[\sigma]$ does not necessarily imply the maximization of $I_{x \to y}$. In the next, we show how consideration of a generalized entropy production, allows us to overcome this problem.

Generalized Fluctuation Theorem : We prove the following inequality below:

$$I_{x \to y} \geq -E[\Theta_\nu^t]. \tag{3}$$

Here, we define the following generalized forms of the entropy production in terms of a conditional distribution ν:

$$\Theta_\nu^t \equiv \log \frac{\pi(x^{t+1}|y^{t+1})}{\nu(x^t|x^{t+1}, y^t)}, \quad \widehat{\Theta}_\nu^t \equiv \Theta_\nu^t + \log \frac{p_s(x^t|y^t)}{p_s(x^{t+1}|y^{t+1})}. \tag{4}$$

We can regard ν as representing physical quantities computed in the neural system on the basis of x^t, y^t, y^{t+1} and adjusted through learning (see supplemental materials). First, we have the apparent identity

$$\exp\left[\log p_s(x^t|y^t) + \log \pi(x^{t+1}|y^{t+1}) - \widehat{\Theta}_\nu^t\right]$$
$$= \exp\left[\log p_s(x^{t+1}|y^{t+1}) + \log \nu(x^t|x^{t+1}, y^t)\right]. \tag{5}$$

Multiplying both sides by $p_s(y^{t+1}, y^t)$ and summing them over relevant random variables, we obtain a generalized form of the fluctuation theorem:

$$E\left[e^{-I_{tr}^{t+1} - \widehat{\Theta}_\nu^t}\right] = 1, \quad I_{tr}^{t+1} \equiv \log \frac{p_s(y^{t+1}|x^t, y^t)}{p_s(y^{t+1}|y^t)}. \tag{6}$$

Applying Jensen's inequality ($\exp E[F(Z)] < E[\exp F(Z)]$, which applies to any random variable Z and any function F) to Eq.(6) gives

$$E\left[-I_{tr}^{t+1} - \widehat{\Theta}_\nu^t\right] \leq 0. \tag{7}$$

Noting that $E[\widehat{\Theta}_\nu^t] = E[\Theta_\nu^t]$ and $I_{x \to y} = E[I_{tr}^{t+1}]$, we have the inequality in Eq.(3). It is found that, for fixed π, the right-hand side of Eq.(3) is maximal if and only if

$$\nu(x^t|x^{t+1}, y^t) = p_s(x^t|x^{t+1}, y^t). \tag{8}$$

Furthermore, we can prove that the equality in Eq.(3) holds if and only if, in addition to Eq.(8), the mutual information takes the maximal value for the fixed π and hence satisfies $I[x^t; y^t] = H[y^t]$, under suitable conditions (see supplemental materials). If the neural network has sufficient capacity, it is expected that there is some optimal π that maximizes both $I_{x \to y}$ and $I[x^t; y^t]$, simultaneously. In this case, the above analysis implies that the optimal π is obtained by maximizing $-E[\Theta_\nu^t]$ with respect to π and ν. In conclusion, we find that, for a neural network with a large capacity, the maximization of $-E[\Theta_\nu^t]$ leads to the maximization of $I_{x \to y}$ and $I[x^t; y^t]$. Because the maximization of $I[x^t; y^t]$ served as the definition of Infomax in previous studies [1–3], the maximization of $-E[\Theta_\nu^t]$ provides a generalized Infomax.

Structures of Neural Networks : To maximize $-E[\Theta_\nu^t]$, the neural network must be able to adjust π and ν to optimal conditional distributions through learning.

For this purpose, in the remainder of this Letter, we parameterize $\pi(x^t|y^t)$ as

$$\pi(x^t|y^t) = \prod_{\ell=1}^{L} \prod_{i=N_{\ell-1}+1}^{N_\ell} \pi_i(x_i^t|y^t, \{x_k^t\}_{k=1}^{N_{\ell-1}}), \qquad (9)$$

$$\pi_i(x_i^t = 1|y^t, \{x_k^t\}_{k=1}^{N_{\ell-1}}) = g(e_i^t),$$

$$e_i^t = \sum_{1 \le j \le M_\ell} \rho_{ij}\{g(\xi_{(\ell),j}^t) - \frac{1}{2}\} - h_0,$$

$$\xi_{(\ell),j}^t = \sum_{1 \le k \le d} v_{jk}^{(\ell)} y_k^t + \sum_{1 \le k \le N_{\ell-1}} w_{jk}^{(\ell)} x_k^t - h_j^{(\ell)}. \quad (10)$$

Here, $g(e_i^t)$ is the logistic function $(1 + \tanh(e_i^t))/2$. Equation(10) describes the situation in which each neuron computes its own transition probability, π_i, through the intermediate units $\xi_{(\ell),j}^t$, with the adjustable parameters $\rho_{ij}, v_{jk}^{(\ell)}, w_{jk}^{(\ell)}$ and $h_j^{(\ell)}$ and the constant parameter h_0. These parameters represent the synaptic strengths and intrinsic properties of the neurons and intermediate units. Note that we assume a layered structure of the system, as illustrated in Fig.2, in which neuron x_i^t in layer ℓ receives an input $g(e_i^t)$ from the neurons in layers 1 through $\ell - 1$ and the environment through the ℓ-th intermediate layer. It is known that an arbitrary continuous mapping of $\{y_k^t\}_{k=1}^{d}$ and $\{x_k^t\}_{k=1}^{N_{\ell-1}}$ to e_i^t can be approximated by the last two lines in Eq.(10) to arbitrary precision if the number of the intermediate units, M_ℓ, is sufficiently large [19]. Thus, any conditional probability of the form given in Eq.(9) can be represented in terms of $g(e_i^t)$, as in Eq.(10). Increasing the number of layers of the neural network increases its capability to represent various conditional probabilities. We believe that the capability to represent a wide variety of conditional probabilities will allow for realization of the optimal π, and therefore such capability is necessary for our purposes. We model ν in the same way as π. Note that ν is not used for the realization of neural states. We consider the situation in which only the values of $\log \nu(x^t|x^{t+1}, y^t)$ are calculated through some biological mechanisms based on the realized states, x^t, x^{t+1} and y^t (see supplemental materials for details regarding ν).

A Simple Model of Animals Learning to Explore Environments : We have shown that neural systems can maximize the transfer entropy and mutual information through a learning mechanism based on a generalized fluctuation theorem. In order to characterize the present learning mechanism, we must clarify the role of the maximization of the transfer entropy in biological contexts, while that of the mutual information has been investigated in previous studies [1–3]. In the following sections, we show that the maximization of the transfer entropy can be understood as a mechanism for the active exploration by an animal of its environment. In order to
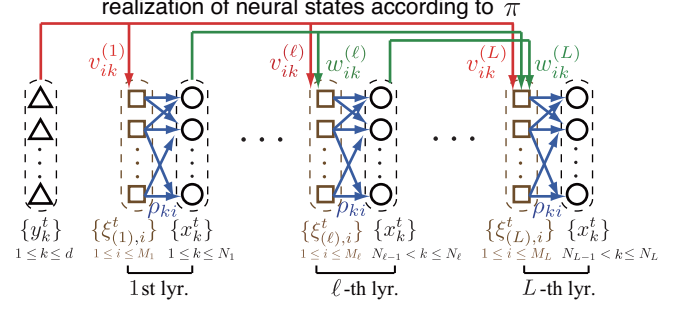

FIG. 2. (Color online). An illustration of the layered neural networks for the modeling of π (ν is modeled in a similar manner).

clearly demonstrate this effect in biological contexts, we introduce a learning problem in which an animal seeks to obtain rewards (e.g., food, water, etc.) through active exploration.

Concretely, an animal with a neural system represented by the state x^t moves around in a two-dimensional grid. At each position in the grid, a value of a reward associated with that position is defined (Fig.3(a)). Specifically, in each timestep, the animal takes either one step or zero steps, with the number and direction determined by the values of the specialized neurons, as shown in Fig.3(b). The state of the environment, y^t, is specified by the position of the animal and the status of reward configuration in the grid. At each timestep, the animal "receives" the reward $r^t = r(y^t)$ at its present position. As shown in Fig.3(a), at most positions in the grid, the reward takes a negative value fixed throughout the simulation. Such a negative reward is interpreted as a punishment. The size of the punishment is minimal in the center of the grid and increases in each direction moving away from the center. At eight (fixed) positions in the outer region of the grid, there are positive rewards. The value of each is initially R. If such a positive reward is visited by the animal, the reward at the position is 0 for the subsequent 100 timesteps and then reset to R. The animal receives inputs from the environment as twelve real variables $\{y_k^t\}$ ($1 \le k \le 12$). The inputs $\{y_k^t\}$ consist of the coordinate values of the animal's position in the grid ($k = 1, 2$), the presence or absence of a reward R at the animal's current position ($y_3^t = 1$ or 0) and the values of the rewards at all positions within one step of the current position ($4 \le k \le 12$), as shown in Fig.3(c). This set of values allows the animal to predict the immediate consequence of its movement. Initially, the model parameters that determine π are set in such a way that the animal primarily attempts to avoid negative rewards, mimicking the innate behavior of real animals (see supplemental materials). With this model, it is very natural to consider maximization of
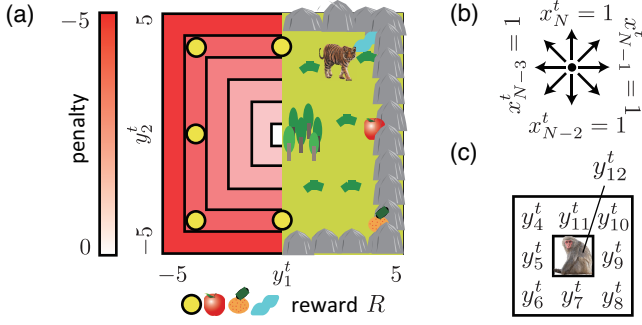
FIG. 3. (Color). A simple model of an animal exploring in a two-dimensional grid: (a) the configuration of reward in the grid, (b) animal's movement according to the values of the specialized neurons, (c) surrounding reward values as input variables.

the average reward, $\mathrm{E}[r^t]$, by adjusting the animal's behavior represented by π, because animals must do so for survival. This maximization problem is called a *reinforcement learning problem*. However, in general, it is known that algorithms that simply maximize $\mathrm{E}[r^t]$ do not reach an optimal outcome in most realistic situations because there is a lack of new experiences, unless some mechanisms for active exploration are included [20, 21]. In the present case, in order to obtain the rewards R to realize a larger $\mathrm{E}[r^t]$, the animal must possess a mechanism that allows it to explore the outer region and tolerate the punishment incurred there. In the following, we show that maximization of the transfer entropy in addition to the average reward provides this mechanism.

We consider the following learning problem:

$$\max \; \left(\mathrm{I}_{x\to y} + \beta \mathrm{E}[r^t]\right), \quad (11)$$

where $\beta \geq 0$. First, we theoretically analyze the optimal π for the above problem. Since the neural control over the environment is deterministic in the above model; that is, for given x^t and y^t, we have $\mu(y^{t+1}|x^t, y^t) = 1$ or 0, the optimization of $\pi(x^t|y^t)$ reduces to that of $\alpha(y^{t+1}|y^t) \equiv \sum_{x^t} \mu(y^{t+1}|x^t, y^t)\pi(x^t|y^t)$. As we know from the basic theory of reinforcement learning, it is helpful for analysis of the maximization problem treated here to consider the following functions of $y \in \mathcal{Y}$:

$$V_{r,\alpha}^{(\gamma)}(y) = \mathrm{E}\left[\sum_{s=1}^{\infty} \gamma^s r^{t+s}\Big|y^t = y\right] - \mathrm{E}\left[\sum_{s=1}^{\infty} \gamma^s r^{t+s}\right],$$

$$V_{I,\alpha}^{(\gamma)}(y) = \mathrm{E}\left[\sum_{s=1}^{\infty} \gamma^s I_{tr}^{t+s}\Big|y^t = y\right] - \mathrm{E}\left[\sum_{s=1}^{\infty} \gamma^s I_{tr}^{t+s}\right], \quad (12)$$

where γ is a parameter satisfying $0 \leq \gamma < 1$. The above quantities with $\gamma \to 1$ represent the average amounts of "excess reward" and "excess information", obtained from the initial state y until the system has relaxed into the steady state. This is analogous to the definition of

the "excess heat" in steady-state thermodynamics [22, 23]. With these limits, we can prove that the learning problem, Eq.(11), has a unique optimal distribution α^* of the following form (see supplemental materials):

$$\alpha^*(y^{t+1}|y^t) \propto \exp[\beta\{r^{t+1}+V_{r,\alpha^*}^{(1)}(y^{t+1})\}+V_{I,\alpha^*}^{(1)}(y^{t+1})]. (13)$$

Inspecting Eq.(13), we understand that the animal shows the following three types of behaviors determined by the value of β. In the case with finite $\beta(>0)$, the animal moves with high probability in a direction for which large future reward is expected, and with small (but non-zero) probability in a direction for which small future reward is expected. It is known that such exploratory behavior, with (infrequent) excursions in directions with low expected payoff, is necessary for neural systems to find larger rewards [20, 21]. Contrastingly, in the case with $\beta \to \infty$, the optimal behavior is deterministic, and exploration is stifled. In the case with $\beta = 0$, the animal is completely insensitive to the values of reward. Hence, we see behavior that is a compromise between the drive to explore and the drive to acquire large rewards, represented by $\mathrm{I}_{x\to y}$ and $\mathrm{E}[r^t]$, respectively.

Numerical Simulations : In order to confirm the theoretical results obtained in the above, we carried out simulations in which we maximized $\mathrm{E}[\beta r^t - \Theta_\nu^t]$ by applying a stochastic gradient algorithm to the model depicted in Fig.3 (see supplemental materials for the algorithms and discussion of its biological counterparts). It is expected that this maximization will result in the maximization in Eq.(11). We first examine the case with $\beta = \infty$, i.e., that in which the animal attempts to maximize $\mathrm{E}[r^t]$ ($R = 600$). In this case, we observe that the environmental entropy, $\mathrm{H}[y^t]$, decreases monotonically and that $\mathrm{E}[r^t]$ becomes fixed at zero (Fig.4(b),(c)). This indicates that the animal has learned only to avoid the outer areas and remains for all times at the origin. Hence, the learning has essentially failed. By contrast, setting $\beta = 0.1$ and $R = 0$, we observe that $-\mathrm{E}[\Theta_\nu^t]$ increases monotonically in Fig.4(a). We also observe that $\mathrm{H}[y^t]$ increases in a similar manner to $-\mathrm{E}[\Theta_\nu^t]$ and that $\mathrm{I}[x^t;y^t]$ almost realizes the maximal value, and satisfies $\mathrm{I}[x^t;y^t] = \mathrm{H}[y^t]$ (Fig.4(b)). Hence, we have confirmed that the maximization of $-\mathrm{E}[\Theta_\nu^t]$ leads to exploration and maximization of $\mathrm{I}[x^t;y^t]$, as theoretically predicted above. Finally, with $\beta = 0.1$ and $R = 600$, we find that the animal is able to increase $\mathrm{E}[r^t]$ through the exploration (Fig.4(c)).

Conclusion : We have shown on the basis of theoretical and numerical analysis that assuming that the learning process exhibited by neural systems is based on a principle described by a generalized fluctuation theorem, this system will learn an effective form of exploring behavior (by maximizing the transfer entropy, $\mathrm{I}_{x\to y}$)
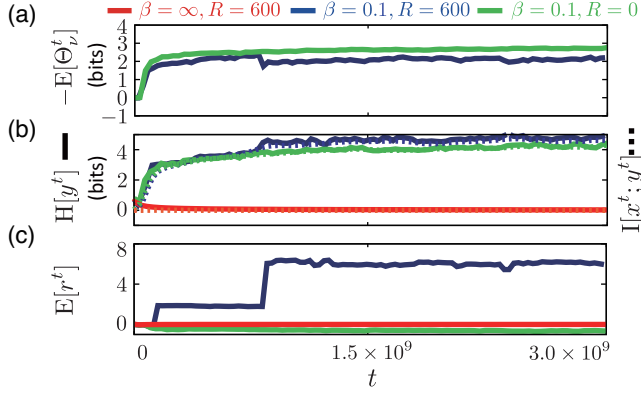
FIG. 4. (Color). The values of (a) $-\mathrm{E}[\Theta_\nu^t]$ (the red line is out of range), (b) $\mathrm{H}[y^t]$ and $\mathrm{I}[x^t; y^t]$, and (c) $\mathrm{E}[r^t]$, during the course of learning. Neural networks with $L = 4$, $N_1 = 30$, $N_2 = 60$, $N_3 = 62$, $N_4 = 64$, $M_1 = M_2 = 120$, and $M_3 = M_4 = 60$ were simulated with $\beta = \infty, 0.1$, and $R = 0, 600$.

and acquiring information about its environment (by maximizing the mutual information, $\mathrm{I}[x^t; y^t]$). Although informational quantities other than the transfer entropy have been considered as mechanisms for the exploration [24–28], it has not been elucidated how those quantities are maximized in neural systems. We believe that use of the transfer entropy as a mechanism for exploration is more plausible, because the present learning mechanism can be utilized for it. Although the demonstration is limited to the case of Markovian environmental dynamics and neural networks without memory, this work will be generalized to more complex systems in the near future using the foundation laid by the present work.

Supplemental Materials

Proof of The Maximization of The Mutual Information, $\mathrm{I}[x^t; y^t]$: In this section, we prove that the maximization of the mutual information, $\mathrm{I}[x^t; y^t]$, and Eq.(8) are equivalent to equality in Eq.(3), under suitable conditions. First, we note that we can replace $\nu(x^t|x^{t+1}, y^t)$ by $\nu(x^t|x^{t+1}, y^t, y^{t+1})$ in Eqs.(3), (4), (5), (6) and (7). In this case, equality in Eq.(3) follows from equality in the Jensen inequality ($\exp F(Z) = \mathrm{E}[\exp F(Z)]$ with probability 1):

$$e^{-I_{tr}^{t+1} - \widehat{\Theta}_\nu^t} = \mathrm{E}[e^{-I_{tr}^{t+1} - \widehat{\Theta}_\nu^t}] = 1. \tag{14}$$

This implies $-I_{tr}^{t+1} - \widehat{\Theta}_\nu^t = 0$ with probability 1. By rearranging terms, we have

$$\nu(x^t|x^{t+1}, y^t, y^{t+1}) = p_s(x^t|y^{t+1}, y^t, x^{t+1}). \tag{15}$$

Hence, by including y^{t+1} as a conditioning variable of ν, we can easily obtain the equality. However, by reducing the number of conditioning variables, we can also obtain the maximization of the mutual information, $\mathrm{I}[x^t; y^t]$, as we noted in the main text. We prove this in the following.

First, we obtain an explicit expression of the optimal $\nu(x^t|x^{t+1}, y^t)$ in Eq.(8) from the following inequality:

$$\mathrm{E}\left[\log \frac{\nu(x^t|y^t, x^{t+1})}{p_s(x^t|y^t, x^{t+1})}\right] = \sum_{y^t, x^{t+1}} p_s(y^t, x^{t+1}) \sum_{x^t} p_s(x^t|y^t, x^{t+1}) \log \frac{\nu(x^t|y^t, x^{t+1})}{p_s(x^t|y^t, x^{t+1})} \leq 0. \tag{16}$$

The above inequality is derived from the inequality $F - 1 \geq \log F$ for positive real F, and thus, the optimality condition in Eq.(8) is obtained from the equality, $F - 1 = \log F \leftrightarrow F = 1$ with probability 1:

$$\nu(x^t|x^{t+1}, y^t) = p_s(x^t|y^t, x^{t+1}). \tag{17}$$

Then, in order to analyze the equality condition of Eq.(3) for $\nu(x^t|x^{t+1}, y^t)$, we calculate the difference Δ between the values of $-\mathrm{E}[\Theta_\nu^t]$ as calculated with Eqs. (17) and (15), writing

$$\Delta = \mathrm{E}\left[\log \frac{p_s(x^t|y^t, y^{t+1}, x^{t+1})}{p_s(x^t|y^t, x^{t+1})}\right] = \mathrm{I}[y^{t+1}; x^t|x^{t+1}, y^t]. \tag{18}$$

Because $-\mathrm{E}[\Theta_\nu^t] = \mathrm{I}_{x \to y}$ in the case with Eq.(15), we obtain

$$\mathrm{I}_{x \to y} + \mathrm{E}[\Theta_\nu^t] = \mathrm{I}[y^{t+1}; x^t|x^{t+1}, y^t], \tag{19}$$

in the case with Eq.(17).

Next, we show that the condition

$$\mathrm{I}[y^{t+1}; x^t|x^{t+1}, y^t] = 0, \tag{20}$$

is equivalent to the maximization of the mutual information,

$$\mathrm{I}[x^t; y^t] = \mathrm{H}[y^t], \tag{21}$$

on the assumption that the environmental state space $\mathcal{Y}$ is not too finely partitioned in comparison with the precision of neural control over the environment. Precisely, we assume that there is no coarse-grained partition $\mathcal{Y}'$ of $\mathcal{Y}$ such that the neural control has the same precision on the two partitions $\mathcal{Y}$ and $\mathcal{Y}'$ of the environmental state space. We also assume that $p_s(y^{t+1}|y^t) \neq 1$ for all y^t and y^{t+1}, which holds for most sets of values of the model parameters in a general model. A coarse-grained partition $\mathcal{Y}'$ is a set of subcollections of $\mathcal{Y}$ such that $y' \cap y'' = \emptyset$ for any $y' \neq y'' \in \mathcal{Y}'$ and $\cup_{y' \in \mathcal{Y}'} y' = \mathcal{Y}$. For any such coarse-grained partition $\mathcal{Y}'$, we require

$$\mathrm{I}[y^{t+1,'}; x^t|y^t] \neq \mathrm{I}[y^{t+1}; x^t|y^t], \ y^{t+1} \in y^{t+1,'} \in \mathcal{Y}', \tag{22}$$

where we have defined the random variable $y^{t+1,'}$, which takes values in $\mathcal{Y}'$ with $y^{t+1} \in y^{t+1,'}$. Under this assumption, we first show that the conditional mutual information,

$$\mathrm{I}[y^{t+1}; x^{t+1}|y^t] = \sum_{x^{t+1}, y^{t+1}, y^t} p_s(y^t, y^{t+1}, x^{t+1}) \log \frac{p_s(y^{t+1}|x^{t+1}, y^t)}{p_s(y^{t+1}|y^t)}, \tag{23}$$

must be maximal. Note that the conditional mutual information takes its maximal value and hence satisfies

$$I[x^{t+1}; y^{t+1}|y^t] = H[y^{t+1}|y^t] \tag{24}$$

if and only if $p_s(y^{t+1}|x^{t+1}, y^t) = 1$ with probability 1. Thus, to obtain the desired result, we show that $p_s(y^{t+1}|x^{t+1}, y^t) > 0$ for multiple y^{t+1} with some $y^t = y_0$ and $x^{t+1} = x_0$ contradicts Eq.(20).

First, we define the set

$$\overline{y} = \{y^{t+1} \in \mathcal{Y}|p_s(y^{t+1}|x^{t+1} = x_0, y^t = y_0) > 0\}, \tag{25}$$

and a coarse-grained partition of $\mathcal{Y}$ as

$$\mathcal{Y}' = \{\{y\}\}_{y \in \mathcal{Y} \setminus \overline{y}} \cup \{\overline{y}\}. \tag{26}$$

Here, $\mathcal{Y} \setminus \overline{y}$ is the relative complement of $\overline{y}$ in $\mathcal{Y}$, which consists of all the elements of $\mathcal{Y}$ that are not contained in $\overline{y}$. The assumption in Eq.(22) requires

$$\begin{aligned}
I[y^{t+1}; x^t|y^t] - I[y^{t+1,\prime}; x^t|y^t] &= I[y^{t+1}, y^{t+1,\prime}; x^t|y^t] - I[y^{t+1,\prime}; x^t|y^t] \\
&= I[y^{t+1}; x^t|y^{t+1,\prime}, y^t] \\
&> 0,
\end{aligned} \tag{27}$$

where the first equality holds because y^{t+1} uniquely determines $y^{t+1,\prime}$ and thus the additional inclusion of $y^{t+1,\prime}$ in the first term does not affect the value of the conditional mutual information. Also, Eq.(20) implies

$$I[y^{t+1}; x^t|y^t, x^{t+1}] = I[y^{t+1}; x^t|y^{t+1,\prime}, y^t, x^{t+1}] = 0. \tag{28}$$

Now, recall that the inclusion of additional conditioning variables (in this case, $y^{t+1,\prime}$) always reduces the value of the conditional mutual information. The right-hand side of the above equation can be written as

$$\begin{aligned}
&\mathrm{E}\left[\log \frac{p_s(y^{t+1}|x^t, y^{t+1,\prime}, y^t, x^{t+1})}{p_s(y^{t+1}|y^{t+1,\prime}, y^t, x^{t+1})}\right] \\
&= \mathrm{E}\left[\log\left\{\frac{p_s(x^{t+1}|y^{t+1})p_s(y^{t+1}|y^{t+1,\prime}, y^t, x^t)}{\sum_{\widetilde{y}^{t+1} \in \mathcal{Y}} p_s(x^{t+1}|\widetilde{y}^{t+1})p_s(\widetilde{y}^{t+1}|y^{t+1,\prime}, y^t, x^t)} \frac{\sum_{\widehat{y}^{t+1} \in \mathcal{Y}} p_s(x^{t+1}|\widehat{y}^{t+1})p_s(\widehat{y}^{t+1}|y^{t+1,\prime}, y^t)}{p_s(x^{t+1}|y^{t+1})p_s(y^{t+1}|y^{t+1,\prime}, y^t)}\right\}\right] \\
&= 0,
\end{aligned} \tag{29}$$

with the dummy variables $\widetilde{y}^{t+1}$ and $\widehat{y}^{t+1}$ having the same (conditional) distributions as y^{t+1}. The above equality requires that the argument of the logarithm be 1 with probability 1, since $F - 1 \geq \log F$, $F - 1 = \log F \leftrightarrow F = 1$ and

$$\begin{aligned}
&-\mathrm{E}\left[\log \frac{p_s(y^{t+1}|x^t, y^{t+1,\prime}, y^t, x^{t+1})}{p_s(y^{t+1}|y^{t+1,\prime}, y^t, x^{t+1})}\right] \\
&= \sum_{x^t, y^{t+1,\prime}, y^t, x^{t+1}} p_s(x^t, y^{t+1,\prime}, y^t, x^{t+1}) \sum_{y^{t+1}} p_s(y^{t+1}|x^t, y^{t+1,\prime}, y^t, x^{t+1}) \log \frac{p_s(y^{t+1}|y^{t+1,\prime}, y^t, x^{t+1})}{p_s(y^{t+1}|x^t, y^{t+1,\prime}, y^t, x^{t+1})} \\
&\leq \sum_{x^t, y^{t+1,\prime}, y^t, x^{t+1}} p_s(x^t, y^{t+1,\prime}, y^t, x^{t+1}) \sum_{y^{t+1}} p_s(y^{t+1}|x^t, y^{t+1,\prime}, y^t, x^{t+1}) \left(\frac{p_s(y^{t+1}|y^{t+1,\prime}, y^t, x^{t+1})}{p_s(y^{t+1}|x^t, y^{t+1,\prime}, y^t, x^{t+1})} - 1\right) \\
&= 0.
\end{aligned} \tag{30}$$

Hence, noting that $p_s(x^{t+1} = x_0|y^{t+1}) > 0$ for $y^{t+1} \in \overline{y}$, we have

$$\frac{p_s(y^{t+1}|y^{t+1,\prime} = \overline{y}, y^t, x^t)}{p_s(y^{t+1}|y^{t+1,\prime} = \overline{y}, y^t)} = \frac{\sum_{\widehat{y}^{t+1} \in \mathcal{Y}} p_s(x^{t+1} = x_0|\widehat{y}^{t+1})p_s(\widehat{y}^{t+1}|y^{t+1,\prime} = \overline{y}, y^t)}{\sum_{\widetilde{y}^{t+1} \in \mathcal{Y}} p_s(x^{t+1} = x_0|\widetilde{y}^{t+1})p_s(\widetilde{y}^{t+1}|y^{t+1,\prime} = \overline{y}, y^t, x^t)}, \quad \forall y^{t+1} \in \overline{y}. \tag{31}$$

Furthermore, Eq.(31) with Eq.(27) implies

$$\frac{p_s(y^{t+1}|y^{t+1,\prime} = \overline{y}, y^t, x^t)}{p_s(y^{t+1}|y^{t+1,\prime} = \overline{y}, y^t)} = c \neq 1, \tag{32}$$

for all $y^{t+1} \in \overline{y}$ and some y^t and x^t, noting

$$\frac{p_s(y^{t+1}|y^{t+1,\prime}, y^t, x^t)}{p_s(y^{t+1}|y^{t+1,\prime}, y^t)} = 1, \quad \forall y^{t+1,\prime} \in \mathcal{Y}' \setminus \{\overline{y}\}. \tag{33}$$
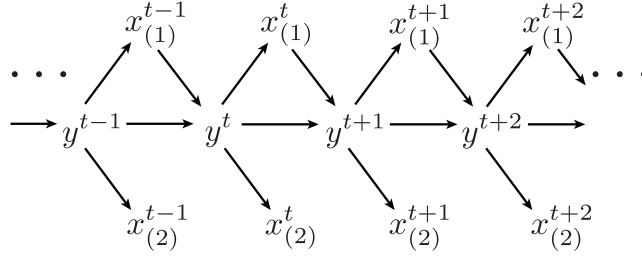
FIG. 5. Representation of the dynamics with an additional neural network z^t as a causal network.

Here, note that $c = 1$ in Eq.(32) with Eq.(33) implies $I[y^{t+1}; x^t | y^{t+1,\prime}, y^t] = 0$, violating the assumption in Eq.(27). However, this implies

$$1 = \sum_{y^{t+1} \in \overline{y}} p_s(y^{t+1} | y^{t+1,\prime} = \overline{y}, y^t, x^t) = c \sum_{y^{t+1} \in \overline{y}} p_s(y^{t+1} | y^{t+1,\prime} = \overline{y}, y^t) = c \neq 1. \tag{34}$$

This contradiction completes the proof of the maximality of the conditional mutual information, Eq.(24).

Next, we show the equivalence of the maximality of the conditional mutual information and the maximality of the mutual information. As we have discussed, Eq.(24) implies

$$p_s(y^{t+1} | x^{t+1}, y^t) = \frac{p_s(y^t | y^{t+1}) p_s(y^{t+1} | x^{t+1})}{\sum_{\widetilde{y}^{t+1} \in \mathcal{Y}} p_s(y^t | \widetilde{y}^{t+1}) p_s(\widetilde{y}^{t+1} | x^{t+1})} = 1, \tag{35}$$

with probability 1. Then, the assumption $p_s(y^{t+1} | y^t) \neq 1$ implies that $p_s(y^t | y^{t+1})$ is positive for multiple y^{t+1}. Thus, the condition Eq.(35) implies $p_s(y^{t+1} | x^{t+1}) = 1$ with probability 1, or equivalently, that the mutual information, $I[x^t; y^t]$, is maximal and hence satisfies

$$I[x^t; y^t] = H[y^t]. \tag{36}$$

Conversely, the equality in Eq.(36) implies

$$\begin{aligned} I[y^{t+1}; x^t | x^{t+1}, y^t] &= H[y^{t+1} | x^{t+1}, y^t] - H[y^{t+1} | x^t, x^{t+1}, y^t] \\ &\leq H[y^{t+1} | x^{t+1}] \\ &= H[y^t] - I[x^t; y^t] \\ &= 0. \end{aligned} \tag{37}$$

Thus we have recovered the condition Eq.(20). This completes the proof that equality in Eq.(3) is equivalent to Eq.(8) and Eq.(21).

In the above proof, we note that different definitions of ν also lead to the maximization of the mutual information $I[x^t; y^t]$ in different manners, although we do not note this point in the main text for the simplicity of presentation. Concretely, we consider the model described by the causal network in Fig.5 by splitting x^t into $x^t_{(1)}$ and $x^t_{(2)}$, and define a generalized entropy production as

$$\Theta^t_\nu = \log \frac{\pi(x^{t+1}_{(1)} | y^{t+1})}{\nu(x^t_{(1)} | x^t_{(2)}, y^{t+1})}. \tag{38}$$

Then, in the same manner as above, we can prove that the equality, $-E[\Theta^t_\nu] = I_{x_{(1)} \to y}$, implies the maximization of the mutual information, $I[x^t_{(2)}; y^{t,\prime}] = H[y^{t,\prime}]$, in some coarse-grained partition $\mathcal{Y}'$ that satisfies

$$I[x^t_{(1)}; y^{t,\prime} | y^{t+1}] = I[x^t_{(1)}; y^t | y^{t+1}]. \tag{39}$$

Further results in this direction will be investigated in the future reports.

Modeling of ν with a Neural Network : We compute ν in the same way as π, explicitly writing

$$\nu(x^t|x^{t+1}, y^t) = \prod_{\ell=1}^{L} \prod_{i=N_{\ell-1}+1}^{N_\ell} \nu_i(x_i^t|\{x_k^t\}_{k=1}^{N_{\ell-1}}, x^{t+1}, y^t),$$

$$\nu_i(x_i^t = 1|\{x_k^t\}_{k=1}^{N_{\ell-1}}, x^{t+1}, y^t) = g(f_i^{t+1}),$$

$$f_i^{t+1} = \sum_{1 \le j \le M_\ell} \kappa_{ij}\{g(\eta_{(\ell),j}^{t+1}) - \frac{1}{2}\} - m_0,$$

$$\eta_{(\ell),j}^{t+1} = \sum_{1 \le k \le N} z_{jk}^{(\ell)} x_k^{t+1} + \sum_{1 \le k \le N_{\ell-1}} u_{jk}^{(\ell)} x_k^t + \sum_{1 \le k \le d} q_{jk}^{(\ell)} y_k^t - m_j^{(\ell)}. \tag{40}$$

Here, the neurons in the ℓ-th layer receive inputs from $\{x_k^t\}_{1 \le k \le N_{\ell-1}}$, x^{t+1} and y^t through the intermediate units, $\eta_{(\ell),j}^{t+1}$, with the adjustable parameters κ_{ij}, $z_{jk}^{(\ell)}$, $u_{jk}^{(\ell)}$, $q_{jk}^{(\ell)}$ and $m_j^{(\ell)}$ and the constant parameter m_0. This computation may seem strange, because here the neurons receive inputs from the future states. However, this is not problematic, because the goal of the computation is not to realize the states of the neural network but to calculate the value of $\log \nu(x^t|x^{t+1}, y^t)$. Consider the following situation for this computation, for example. The intermediate units in the ℓ-th layer receive inputs at time $t+1$ from x^{t+1} and also from $\{x_k^t\}_{1 \le k \le N_{\ell-1}}$ and y^t through some time-delay mechanisms. These intermediate units send outputs $\{g(\eta_{(\ell),j}^{t+1})\}_{1 \le j \le M_\ell}$ to the neurons in the ℓ-th layer. At this time, the i-th neuron in this layer possesses memory of its own state at time t, x_i^t, through some mechanism. Then, the i-th neuron can compute the value of $\nu_i(x_i^t|\{x_k^t\}_{k=1}^{N_{\ell-1}}, x^{t+1}, y^t)$ as a function of x_i^t and $\{g(\eta_{(\ell),j}^{t+1})\}_{1 \le j \le M_\ell}$. The value of $\log \nu$ is the sum of these values of ν_i over the neurons in the neural network.

Proofs of the Relations Used in the Theoretical Analysis of the Reinforcement Learning Problem : In this section, our goal is to prove Eq.(13). First, we define the following functions called "value functions" in the field of reinforcement learning:

$$\widehat{V}_{r,\alpha}^{(\gamma)}(y) = E\left[\sum_{s=1}^{\infty} \gamma^s r^{t+s} \Big| y^t = y\right],$$

$$\widehat{V}_{I,\alpha}^{(\gamma)}(y) = E\left[\sum_{s=1}^{\infty} \gamma^s I_{tr}^{t+s} \Big| y^t = y\right],$$

$$\widehat{V}_{\alpha}^{(\gamma)}(y) = \widehat{V}_{I,\alpha}^{(\gamma)}(y) + \beta \widehat{V}_{r,\alpha}^{(\gamma)}(y). \tag{41}$$

Then, we can write the learning problem Eq.(11) in terms of the value functions as

$$I_{x \to y} + \beta E[r^t] = \lim_{\gamma \to 1} (1-\gamma)\widehat{V}_{\alpha}^{(\gamma)}(y), \quad \forall y \in \mathcal{Y}. \tag{42}$$

By definition, the value function satisfies the following recursive relation called the "Bellman equation":

$$\widehat{V}_{\alpha}^{(\gamma)}(y^t) = \sum_{y^{t+1} \in \mathcal{Y}} \gamma \alpha(y^{t+1}|y^t)$$

$$\times \{ \beta r(y^{t+1}) - \log \alpha(y^{t+1}|y^t) + \widehat{V}_{\alpha}^{(\gamma)}(y^{t+1})\}. \tag{43}$$

Next, we show that for fixed γ, it is known that an optimal control $\alpha^*(y^{t+1}|y^t)$ maximizes the value function at all $y \in \mathcal{Y}$, in comparison with suboptimal controls. Explicitly, for any control α, the following inequality holds:

$$\widehat{V}_{\alpha^*}^{(\gamma)}(y) \ge \widehat{V}_{\alpha}^{(\gamma)}(y), \quad \forall y \in \mathcal{Y}. \tag{44}$$

In order to prove Eq.(44), we consider the following operator called a backup operator, operating on functions of the environmental state y^t:

$$B\phi(y^t) = \max_{\alpha(y^{t+1}|y^t)} \sum_{y^{t+1} \in \mathcal{Y}} \alpha(y^{t+1}|y^t)\{\gamma r(y^{t+1}) - \gamma \log \alpha(y^{t+1}|y^t) + \gamma \phi(y^{t+1})\}. \tag{45}$$

We first show that this operation results in contraction in the space of functions of environmental states y^t with respect to max norm:

$$\| \phi \|_\infty = \max_{y^t \in \mathcal{Y}} |\phi(y^t)|. \tag{46}$$

For two functions ϕ_1 and ϕ_2, a fixed $\alpha(y^{t+1}|y^t)$, and an operator B_α defined as

$$B_\alpha \phi(y^t) = \sum_{y^{t+1} \in \mathcal{Y}} \alpha(y^{t+1}|y^t)\{\gamma r(y^{t+1}) - \gamma \log \alpha(y^{t+1}|y^t) + \gamma \phi(y^{t+1})\}, \tag{47}$$

we have

$$\| B_\alpha \phi_1 - B_\alpha \phi_2 \|_\infty = \| \sum_{y^{t+1}} \alpha(y^{t+1}|y^t)\gamma\{\phi_1(y^{t+1}) - \phi_2(y^{t+1})\} \|_\infty$$

$$\leq \max_{y^t \in \mathcal{Y}} \left\{ \sum_{y^{t+1}} |\alpha(y^{t+1}|y^t)| \right\} \| \gamma(\phi_1 - \phi_2) \|_\infty$$

$$= \gamma \| \phi_1 - \phi_2 \|_\infty . \tag{48}$$

Then, with the distribution $\alpha^{(i),*}(y^{t+1}|y^t)$ maximizing $B_\alpha \phi_i(y^t)$ $(i = 1, 2)$, we have

$$\| B\phi_1 - B\phi_2 \|_\infty = \| B_{\alpha^{(1),*}}\phi_1 - B_{\alpha^{(2),*}}\phi_2 \|_\infty$$

$$\leq \max_i \| B_{\alpha^{(i),*}}\phi_1 - B_{\alpha^{(i),*}}\phi_2 \|_\infty$$

$$\leq \gamma \| \phi_1 - \phi_2 \|_\infty . \tag{49}$$

This proves that the backup operation yields a contraction of the space of functions on the environmental state space $\mathcal{Y}$, and that there is a unique fixed point of this operation in this space of functions. Because the backup operation always increases the values of any value function at any point in $\mathcal{Y}$, we have Eq.(44). Hence, when we consider the maximality condition of $\widehat{V}_\alpha^{(\gamma)}(y)$ with respect to $\alpha(y^{t+1}|y^t)$, it is sufficient to consider the stationarity condition of $B\widehat{V}_\alpha^{(\gamma)}(y^t)$ by differentiating it with respect to $\alpha(y^{t+1}|y^t)$ and simply putting the derivative of $\widehat{V}_\alpha^{(\gamma)}(y^{t+1})$ to be zero. Solving the stationarity condition with the Lagrange multiplier corresponding to $\sum_{y^{t+1}} \alpha(y^{t+1}|y^t) = 1$, we obtain

$$\alpha^*(y^{t+1}|y^t) \propto \exp[\beta\{r(y^{t+1}) + \widehat{V}_{r,\alpha^*}^{(\gamma)}(y^{t+1})\} + \widehat{V}_{I,\alpha^*}^{(\gamma)}(y^{t+1})]. \tag{50}$$

The optimal condition for the learning problem, the maximization of Eq.(42), is obtained by taking the limit $\gamma \to 1$ in Eq.(50). In order to avoid divergence, we need to replace the value functions in Eq.(50) with the functions defined in Eq.(12) that represent the "excess reward" and "excess information".

Derivation and Biological Plausibility of the Learning Rule : In the gradient ascent method used for the simulation, we update each parameter $\theta \in \{\rho_{ij}, v_{jk}^{(\ell)}, w_{jk}^{(\ell)}, h_j^{(\ell)}, \kappa_{ij}, z_{jk}^{(\ell)}, u_{jk}^{(\ell)}, q_{jk}^{(\ell)}, m_j^{(\ell)}\}$ as follows:

$$\theta^{t+1} = \theta^t + \epsilon \left\{ \tau(\beta r^{t+1} - \Theta_\nu^t)\psi_\theta^t - \frac{\partial}{\partial\theta}\Theta_\nu^t \right\}$$

$$\psi_\theta^t = \frac{1}{\tau}\frac{\partial}{\partial\theta}\log\pi(x^{t+1}|y^{t+1}) + \left(1 - \frac{1}{\tau}\right)\psi_\theta^{t-1}. \tag{51}$$

Here, the constant τ is a positive real number that is large compared with the mixing time of the dynamics. We set ϵ to such a small value that the change in the model parameters in each update does not affect the stationarity on a time scale of τ. Then, in the above learning rule, the expectation value of the change in the parameter θ in each update is equal to the gradient of $\mathrm{E}[\beta r^t - \Theta_\nu^t]$ with respect to θ as we show below. Thus, we can regard the learning rule as a stochastic gradient ascent algorithm to maximize $\mathrm{E}[\beta r^t - \Theta_\nu^t]$.

In the gradient ascent method, we must calculate the gradient of the following quantity with respect to θ:

$$\mathrm{E}[\beta r^t - \Theta_\nu^t] = \sum_{x^t, x^{t+1}, y^t, y^{t+1}} p_s(y^t)\pi(x^t|y^t)\mu(y^{t+1}|y^t, x^t)\pi(x^{t+1}|y^{t+1})\{\beta r(y^{t+1}) - \Theta_\nu^t\}.$$

In this calculation, we find that differentiation of the stationary distribution $p_s(y^t)$ is apparently intractable, while differentiation of the other components is easily carried out. We note, however, that we do not need to differentiate

the stationary distribution explicitly, assuming that the stationary distribution is a smooth function of any model parameter θ. In this case, small changes in $p_s(y^{t-\tau})$ for $\tau \gg 1$ vanish at t and $t-1$, and thus terms including the derivatives of $p_s(y^{t-\tau})$ are negligible (see also [29]). Thus, we can compute the gradient as follows:

$$
\frac{\partial}{\partial\theta}\mathrm{E}[\beta r(y^{t+1}) - \Theta_\nu^t]
$$

$$
= \lim_{\tau\to\infty} \sum_{y^{t-\tau},x^{t-\tau},\cdots,y^{t+1},x^{t+1}} p_s(y^{t-\tau})\frac{\partial}{\partial\theta}\left[\pi(x^{t-\tau}|y^{t-\tau})\prod_{s=0}^{\tau}\mu(y^{t-s+1}|x^{t-s},y^{t-s})\pi(x^{t-s+1}|y^{t-s+1})\{\beta r(y^{t+1}) - \Theta_\nu^t\}\right]
$$

$$
= \lim_{\tau\to\infty} \sum_{y^{t-\tau},x^{t-\tau},\cdots,y^{t+1},x^{t+1}} p_s(y^{t-\tau},x^{t-\tau},\cdots,y^{t+1},x^{t+1})\sum_{s=0}^{\tau+1}\frac{\frac{\partial}{\partial\theta}\pi(x^{t-s+1}|y^{t-s+1})}{\pi(x^{t-s+1}|y^{t-s+1})}\{\beta r(y^{t+1}) - \Theta_\nu^t\}
$$

$$
+ \lim_{\tau\to\infty} \sum_{y^{\tau},x^{t-\tau},\cdots,y^{t+1},x^{t+1}} p_s(y^{t-\tau},x^{t-\tau},\cdots,y^{t+1},x^{t+1})\frac{\partial}{\partial\theta}\{\beta r(y^{t+1}) - \Theta_\nu^t\}
$$

$$
= \lim_{\tau\to\infty} \mathrm{E}\left[\{\beta r(y^{t+1}) - \Theta_\nu^t\}\sum_{s=0}^{\tau+1}\frac{\partial}{\partial\theta}\log\pi(x^{t-s+1}|y^{t-s+1})\right] - \mathrm{E}\left[\frac{\partial}{\partial\theta}\Theta_\nu^t\right]. \tag{52}
$$

Note that the third equality follows from $\frac{\partial}{\partial\theta}r(y^{t+1}) = 0$. In order to decompose the expectation values into time-stepwise quantities, we introduce the auxiliary variable ψ_θ^t, defined through

$$
\psi_\theta^t = \frac{1}{\tau}\frac{\partial}{\partial\theta}\log\pi(x^{t+1}|y^{t+1}) + (1 - \frac{1}{\tau})\psi_\theta^{t-1}, \quad \text{and} \quad \psi_\theta^t = 0 \;\; (t \le 0). \tag{53}
$$

Then, we have

$$
\psi_\theta^t = \frac{1}{\tau}\sum_{s=0}^{\infty}(1 - \frac{1}{\tau})^s\frac{\partial}{\partial\theta}\log\pi(x^{t-s+1}|y^{t-s+1}).
$$

If the process under consideration is stationary, ψ_θ^t approaches the long-time average of $\frac{\partial}{\partial\theta}\log\pi(x^{t+1}|y^{t+1})$ as $\tau \to \infty$ and $t/\tau \to \infty$. Similarly, assuming that the correlation of $\beta r(y^{t+1}) - \Theta_\nu^t$ with $\frac{\partial}{\partial\theta}\log\pi(x^{t-\tau+1}|y^{t-\tau+1})$ is small for $\tau \gg 1$ and that $T \gg \tau$, we have

$$
\mathrm{E}\left[\{\beta r(y^{t+1}) - \Theta_\nu^t\}\sum_{s=0}^{\tau+1}\frac{\partial}{\partial\theta}\log\pi(x^{t-s+1}|y^{t-s+1})\right] \approx \frac{\tau}{T}\sum_{t=1}^{T}(\beta r(y^{t+1}) - \Theta_\nu^t)\psi_\theta^t. \tag{54}
$$

Then, applying a well-known argument in stochastic approximation theory [30], we obtain the learning rule given in Eq.(51) as a stepwise approximation of the gradient in Eq.(52).

Finally, we derive the exact form of the learning rule with respect to several θ and present its interpretation. Note that Eq.(51) is composed of $\log\pi(x^{t+1}|y^{t+1})$, $\log\nu(x^t|x^{t+1},y^t)$ and their derivatives with respect to θ. First, we show that these components are easily calculated in a neuron-wise manner. Note that Θ_ν^t, $\log\pi(x^{t+1}|y^{t+1})$ and $\log\nu(x^t|x^{t+1},y^t)$ are decomposed as

$$
\Theta_\nu^t = \log\pi(x^{t+1}|y^{t+1}) - \log\nu(x^t|x^{t+1},y^t)
$$

$$
= \sum_{\ell=1}^{L}\sum_{i=N_{\ell-1}+1}^{N_\ell}\log\pi_i(x_i^{t+1}|y^{t+1},\{x_k^{t+1}\}_{k=1}^{N_{\ell-1}}) - \sum_{\ell=1}^{L}\sum_{i=N_{\ell-1}+1}^{N_\ell}\log\nu_i(x_i^t|\{x_k^t\}_{k=1}^{N_{\ell-1}},x^{t+1},y^t)
$$

$$
= \sum_i\log\chi(e_i^{t+1},x_i^{t+1}) - \sum_i\log\chi(f_i^t,x_i^t), \tag{55}
$$

where

$$
\chi(a,b) = \begin{cases} g(a) & , \quad \text{if} \;\; b = 1, \\ -g(a) & , \quad \text{if} \;\; b = 0. \end{cases}
$$

$$
\tag{56}
$$

Then, the derivatives of Θ^t_ν, $\log \pi(x^{t+1}|y^{t+1})$ and $\log \nu(x^t|x^{t+1}, y^t)$ (with respect to $v^{(\ell)}_{jk}, \rho_{ij}, z^{(\ell)}_{jk}, \kappa_{ij}$, for example) are calculated as follows. First, denoting the derivative of $\chi(a,b)$ with respect to a as $\chi_a(a,b)$,

$$\frac{\partial}{\partial v^{(\ell)}_{jk}} \log \pi(x^{t+1}|y^{t+1}) = \sum_{i=N_{\ell-1}+1}^{N_\ell} \frac{\partial}{\partial v^{(\ell)}_{jk}} \log \pi_i(x^{t+1}_i|y^{t+1}, \{x^{t+1}_k\}^{N_{\ell-1}}_{k=1})$$

$$= \sum_{i=N_{\ell-1}+1}^{N_\ell} \frac{\chi_a(e^{t+1}_i, x^{t+1}_i)}{\chi(e^{t+1}_i, x^{t+1}_i)} \rho_{ij} g'(\xi^{t+1}_{(\ell),j}) y^{t+1}_k. \tag{57}$$

$$\frac{\partial}{\partial \rho_{ij}} \log \pi(x^{t+1}|y^{t+1}) = \frac{\partial}{\partial \rho_{ij}} \log \pi_i(x^{t+1}_i|y^{t+1}, \{x^{t+1}_k\}^{N_{\ell-1}}_{k=1})$$

$$= \frac{\chi_a(e^{t+1}_i, x^{t+1}_i)}{\chi(e^{t+1}_i, x^{t+1}_i)} g(\xi^{t+1}_{(\ell,j)}). \tag{58}$$

$$\frac{\partial}{\partial z^{(\ell)}_{jk}} \log \nu(x^t|x^{t+1}, y^t) = \sum_{i=N_{\ell-1}+1}^{N_\ell} \frac{\partial}{\partial z^{(\ell)}_{jk}} \log \nu_i(x^t_i|\{x^t_k\}^{N_{\ell-1}}_{k=1}, x^{t+1}, y^t)$$

$$= \sum_{i=N_{\ell-1}+1}^{N_\ell} \frac{\chi_a(f^{t+1}_i, x^t_i)}{\chi(f^{t+1}_i, x^t_i)} \kappa^{(\ell)}_{ij} g'(\eta^{t+1}_{(\ell),j}) x^{t+1}_k. \tag{59}$$

$$\frac{\partial}{\partial \kappa_{ij}} \log \nu(x^t|x^{t+1}, y^t) = \frac{\partial}{\partial \kappa_{ij}} \log \nu_i(x^{t+1}_i|\{x^t_k\}^{N_{\ell-1}}_{k=1}, x^{t+1}, y^t)$$

$$= \frac{\chi_a(f^{t+1}_i, x^t_i)}{\chi(f^{t+1}_i, x^t_i)} g(\eta^{t+1}_{(\ell,j)}). \tag{60}$$

It should be noted that calculations of the derivatives involve quantities only for related neurons and intermediate units. For example, the derivative with respect to $v^{(\ell)}_{jk}$ used only information regarding y^{t+1}_k, $\xi^{t+1}_{(\ell),k}$ and $\{\mu^{t+1}_i, x^{t+1}_i, \rho_{ij}\}$ of the i-th neuron to which the j-th intermediate unit is connected. Thus, we can regard the change in the synaptic strength $v^{(\ell)}_{jk}$ as being determined by the local interactions at the synapse on the j-th intermediate unit. Continuing with this line of argument, we can obtain even more realistic forms of learning rules for actual neural systems. However, we do not go into detail here, because the argument becomes quite complicated and is beyond the scope of the current study.

Initial Values of Model Parameters and Values of Learning Parameters Used in the Simulation : In the numerical simulation of our model of learning, we used initial values of the model parameters that results in behavior in which the animal primarily attempts to avoid negative reward, mimicking innate behavior of real animals. We set the values of the model parameters involved in the inputs to the movement-related neurons as shown in Fig.6. A neuron controlling motion in one of four directions receives connections with relatively strong positive weights, ρ_0, from a specialized intermediate unit (for example, from $\xi^t_{(4),1}$ to x^t_N). The intermediate unit receives connections from the environmental variables that take the values of the rewards within one step of the animal's position, y^t_k ($4 \leq k \leq 12$), with the weight-values v_0, $-v_0$ and 0, as illustrated in Fig.6. These initial values of the weight parameters make the neurons controlling motion take a value of 1 when relative amounts of the reward in the corresponding direction are large. We chose the other weight parameters with small random values in accordance with the following: $\rho_{ij} \sim [-0.05 : 0.05]$ ($i \leq N_2$); $v^{(\ell)}_{ij} \sim [-0.05 : 0.05]$ ($\ell = 1, 2$); $\rho_{ij} = 0$ if $N_2 < i \leq N_4 = N$ and $(i,j) \neq (N,1), (N-1,2), (N-2,1), (N-3,2)$; $v^{(\ell)}_{ij} = 0$ ($\ell = 3, 4$ except the red and blue synaptic weights in Fig.6); $w^{(\ell)}_{ij} \sim [-0.05 : 0.05]$ ($\ell = 1, 2$); $w^{(\ell)}_{ij} = 0$ ($\ell = 3, 4$); $h_0 = \log 20$; $h^{(\ell)}_j = 0$; $\kappa_{ij} \sim [-0.05 : 0.05]$; $z^{(\ell)}_{jk} \sim [-0.05 : 0.05]$; $u^{(\ell)}_{jk} \sim [-0.05 : 0.05]$; $q^{(\ell)}_{jk} \sim [-0.05 : 0.05]$; $m^{(\ell)}_j = 0$; $m_0 = 0$.
In the updates of the model parameters according to Eq.(51), we used the following (fixed) values of learning parameters: $\epsilon = 3.0 \times 10^{-5}$; $\tau = 50$.

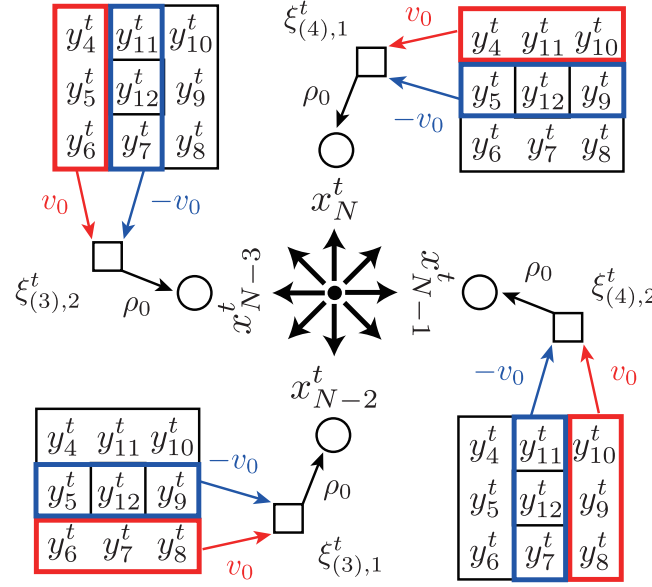[1] R. Linsker, Computer **21**, 105 (1988).

* takashi.hayakawa@riken.jp

FIG. 6. (Color online). Initial values of model parameters for synaptic weights to the movement-related neurons.

[2] A. J. Bell and T. J. Sejnowski, Neural Computation **7**, 1129 (1995).

[3] A. J. Bell and T. J. Sejnowski, Vision Research **37**, 3327 (1997).

[4] T. Tanaka, T. Kaneko, and T. Aoyagi, Neural Computation **21**, 1038 (2009).

[5] T. Hayakawa, T. Kaneko, and T. Aoyagi, Frontiers in Computational Neuroscience **8** (2014).

[6] B. Sallans and G. E. Hinton, The Journal of Machine Learning Research **5**, 1063 (2004).

[7] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, Cognitive Science **9**, 147 (1985).

[8] P. Smolensky, "Parallel distributed processing: explorations in the microstructure of cognition, vol. 1. information processing in dynamical systems: foundations of harmony theory," (1986).

[9] C. Jarzynski, Phys. Rev. Lett. **78**, 2690 (1997).

[10] G. E. Crooks, Phys. Rev. E **60**, 2721 (1999).

[11] D. J. Evans, E. Cohen, and G. P. Morriss, Phys. Rev. Lett. **71**, 2401 (1993).

[12] T. Sagawa and M. Ueda, Phys. Rev. Lett. **104**, 090602 (2010).

[13] U. Seifert, Rep. Prog. Phys. **75**, 126001 (2012).

[14] S. Ito and T. Sagawa, Phys. Rev. Lett. **111**, 180603 (2013).

[15] S. Still, D. A. Sivak, A. J. Bell, and G. E. Crooks, Phys. Rev. Lett. **109**, 120604 (2012).

[16] S. Still, Entropy **16**, 968 (2014).

[17] T. Schreiber, Phys. Rev. Lett. **85**, 461 (2000).

[18] T. M. Cover and J. A. Thomas, *Elements of information theory* (John Wiley & Sons, 2012).

[19] K.-I. Funahashi, Neural Networks **2**, 183 (1989).

[20] R. S. Sutton and A. G. Barto, *Introduction to reinforcement learning*, 1st ed. (MIT Press, Cambridge, MA, USA, 1998).

[21] M. Wiering and M. van Otterlo, *Reinforcement Learning*, Vol. 12 (Springer, 2012).

[22] Y. Oono and M. Paniconi, Prog. of Theor. Phys. Suppl. **130**, 29 (1998).

[23] S.-i. Sasa and H. Tasaki, J. Stat. Phys. **125**, 125 (2006).

[24] M. G. Azar, V. Gómez, and H. J. Kappen, The Journal of Machine Learning Research **13**, 3207 (2012).

[25] J. Peters, K. Mülling, and Y. Altün, in *AAAI* (2010).

[26] S. Still and D. Precup, Theory in Biosciences **131**, 139 (2012).

[27] N. Ay, H. Bernigau, R. Der, and M. Prokopenko, Theory in Biosciences **131**, 161 (2012).

[28] W. Bialek, I. Nemenman, and N. Tishby, Neural Computation **13**, 2409 (2001).

[29] J. Baxter and P. L. Bartlett, *Direct gradient-based reinforcement learning: I. gradient estimation algorithms*, Tech. Rep. (National University, 1999).

[30] H. Robbins and S. Monro, The Annals of Mathematical Statistics **22**, 400 (1951).